Medical Records in Drug and alcohol abuse treatment centers :

Data standards needed to get you there

Clement J. McDonald, M.D.
Director, Lister Hill Center for Biomedical Communications
U.S. National Library of Medicine
National Institutes of Health, HHS
Bethesda, MD

Sept 24, 2010. North Bethesda Marriot

Data Collection

- Medical record systems are like egg cartons without eggs - they have storage slots but no content
- And… they are empty when you first buy them
- The computer does not go out and gather data for you

Data collection costs

- It costs to collect data
  - Pharmacists consume 9 minutes (average) for patient med history needed for medication reconciliation
  - Physician order entry takes from 30 seconds to 2 minutes per order on average across different studies.
- The more granular the coding, the longer the data entry menus, and the longer the data entry time
- Answering discrete questions with menus costs more time than saying what you know (narrative)
  - Recall the mail survey W $2 bill.

Standardizing data helps

- Reduce eliminate data collection costs- (pull from existing systems)
  - Laboratories, pharmacy benefit managers,
- Pool data with others who standardize for research and management purposes
- Stability of content over time

There is an art to standardizing

- It requires thinking about data in flight – a message going from one place to another – not at rest where it becomes rooted in a particular software implementation (Remember software always dies or disappears or becomes outdated but data is "forever")

Two kinds of data

- New kinds of things you wish were structured- but now are free form (paper forms or dictated text)
- Those that are already stored in computers in structured way (not just free text)
  - E.g. Laboratory results, Pharmacy records
- Will take them in order

The crux of standardizing things that are are not already structured

- Requires the same detailed work as needed to build a data collection form or questionnaire–

Questions and answers

- The data collection form boils down to a set of inter related questions and answers
  - To build the form you have to determine the data type of each question –
  - If it is numeric , you also have to specify the units of measure and the absolute range
  - If it is coded, then you have to specify the answer list , explicitly
- If you allow narrative comments , dedicate a explicit place (question) for that
  - You can't scribble on the margins on a computer form

More

- Control data greed
- Define your questions
- Cut them in half
- Determine how long it takes to collect and users tolerance
- Cut again if needed
- Validate them

Don't re-invent the wheel

- Look for validated survey instruments that cover your content needs, They are the best possible data collection forms.
- Next best is a set of questions that have been used on a large scale (in studies or administrative environments –and cover the content)
- The 1st saves you the validation work
- The 1st and 2nd guarantee you have some one to share with.

WHERE TO LOOK FOR EXISTING QUESTIONNAIRES

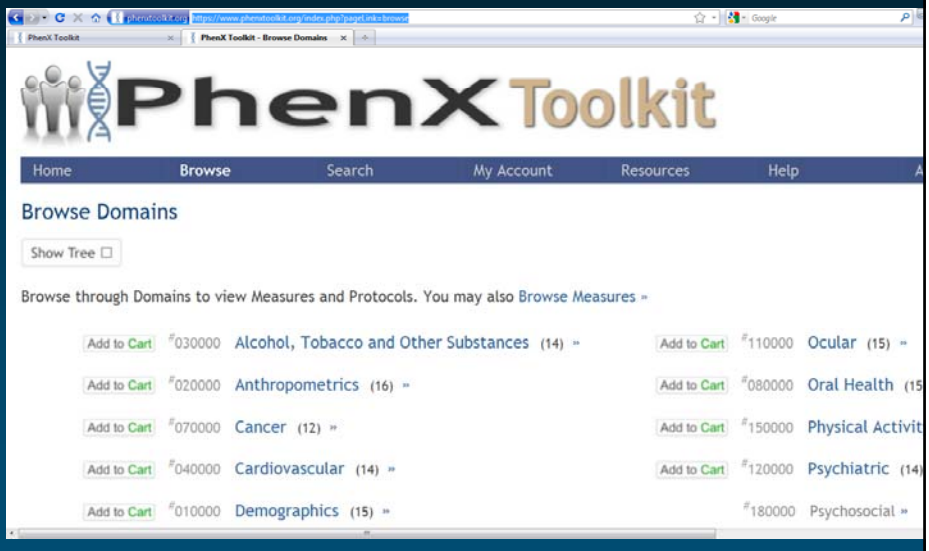
Some places to look

- The literature- (obviously) look for validated surveys
- NHANES – wide spectrum of survey instruments and questions honed over decades
- PROMIS –for variety of functional assessments
- PhenX- broad range of measures for GWAS studies
- Federal Assessment forms (MDS, OASIS, CARE – some parts are reusable -further they *may* become accessible for research purposes

PHENX

- Broad range of formal measures for Genome association and other studies
- All taken from published and formally studies approaches
- Will be tied to LOINC soon
- Includes drug alcohol and substance abuse
- https://www.phenxtoolkit.org/index.php?pageLink=browse

Phenx web site



NCBI's dbGaP

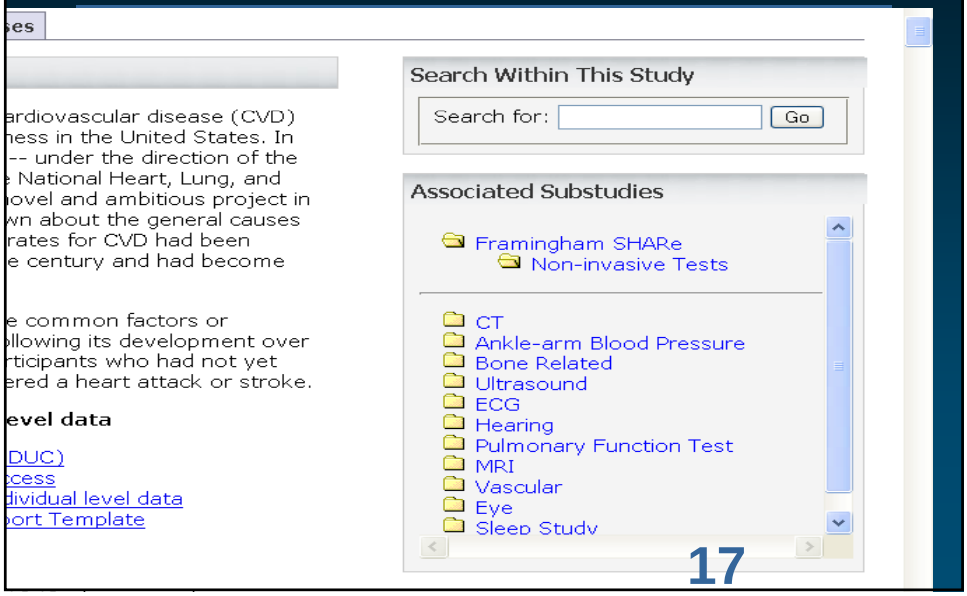
- 100's of longitudinal studies-
- lists all of the questions and the potential answers
- Includes some very large studies (Framingham)
- Good source of question for special disorders

http://www.ncbi.nlm.nih.gov/gap

You can get to much of this data

- Through Db GAP-
  - An NLM-NCBI service
  - http://www.ncbi.nlm.nih.gov/sites/entrez?db=gap
- Includes Framingham and many hundreds of other GWAS (Genome Wide Association Studies)
- Hand out
- You can see the details of what data was collected Down to the exact question and answer menus.
- Can request access to summary data, the patient level data and/or genetic data-

NCBI'S dbGaP -- Framingham



Welcome Trust UK Biobank follow 500K people for 30+years

- It Focus on the 8 commonest diseases
- They have done the work of combing through literature for variables to follow
- See Report of UK population Biomedical Collection Protocol Workshop held at the Royal College of Physicians April 17, 2001 (with Burroughs Welcome support).

E.G collection instruments for DM

- Socio economic variable
- Alcohol use, smoking, exercise
- UK diabetes questionnaire
- Rose angina questionnaire
- Birth weight
- Infectious Hx
- BMI
- Vital signs, step test
- Lots of lab tests

Other Welcome- Collection instruments

- for chronic Pulm disease
  - FEV1 (rather than spirometry and peak flow)
  - MRC breathlessness questionnaire
  - Asthma questionnaire (more than one option)
- Mental health disorders
  - Mini Mental Status
  - "crystallized intelligence" e. g. NART
  - General health questionnaire (GHQ)
  - Depression – BDI or CES-D
  - Lots more

Other subjects that are well covered due to work in technical standards

- Laboratory tests
- Anthropomorphic measurements
- Medications

WHAT IS NEEDED TO PULL EXISTING COMPUTER DATA

Message standards

- Technical standards exist for the carrying data collection instruments
- HL7 version 2.x is king. “Every” hospital and large clinic can deliver Laboratory results, radiology reports, clinician dictation and more via HL7 in almost every hospital and large clinic. Tens of billions of HL7 message are delivered in the US per year.

HL7- the ISO shipping container for results

Stacks of 20' & 40' Shipping Containers

This is what HL7 with a cargo of CBC results looks like

Patient level

PID|||0999999^6^M10||TEST^PATIENT^||1992022
5|F||B|4050 SW WAYWARD BLVD |

Order/report level t

✷OBR|||H9759-0^REG_LAB|24358-4 ^Hemogram^LOINC

✷ Discrete Results

OBX|2|NM||789- 8^RBC^LOINC||4.9|M/mm3| 4.0-5.4..
OBX|3|NM|718-7^HGB^LOINC||12.4|g/dL|12.0- 5.0|..
OBX|4|NM||20570-8^HCT^LOINC||50|%|35-49|H|||F|
OBX|5|NM||30428-7^MCV^LOINC||81|fL|80-94|N|||F|

HL7 sends results in a "table"

- Each discrete result gets its own row
- The yellow column carries the questions
- The Orange column carries the answers
- I used an example with numerically valued answers because if fits on a single slide.
- But it can carry question/answer pairs for questions with multiple choice answers (coded), free text answers, even images as answers

HL7 messages represent a stacked data structure

- Where each value gets its own row.
- This is the rule in EHRs , laboratory systems and pharmacy systems, any where that the number of possible questions can be large.
- It is also what you see in the CDSC result messages and V3 and CDA versions of HL7
- It is different from the flat structure you see most commonly in research- where the column represents the question

Flat structure

Pat ID	Name	surgery date	Hb	DBP	# of BPU	Bypass Minutes	Choles-terol
1234-5	Doe , Jan	12May95	13	95	3	80	180
9999-3	Jones , T	1Aug95	12.5	88	2	90	230
8888-3	Doe Sam	4June95	16	78	0	80	205

Stacked structure

Operational Data Base: One Record Per Observation

Pt ID	Relevant Date	Observation ID	Value	Units	Normal Rang	Place	Observer
Doe J	12-May-95	**Hemoglobin**	13	mg/dl	12.5-15	St Francis	Dr Smith
Doe J	12-May-95	**Hemoglobin**	11.5	mg/dl	12.5-15	St Francis	Dr Smith
Doe J	12-May-95	**Dias BP**	95	mm/Hg	80-140	St Francis	Dr Smith
Doe J	12-May-95	**Dias BP**	110	mm/Hg	80-140	St Francis	Dr Smith
Doe J	13-May-95	**Bypass minutes**	80	min		St Francis	Dr Sleepwell
Doe J	12-May-95	**Diagnosis**	CHF-365 (ICD9)			St Francis	Dr Bloodbank

WHERE DO STANDARD CODES (VOCABULARY) FIT

For clinical observations

- In HL7 and the other message standards, provide the structure and recommend specific codes for specific fields
- For clinical results and orders the key code systems are LOINC (the question) and SNOMED CT (the answer).
- For medication codes Rx.Norm and Rx.terms (that identify drugs and ingredients are the key codes
- All three are required by recent federal regulations

Questions and answers

- In the context of clinical observations
  - LOINC provides standard codes for variables (questions) – esp lab and physical measures and assessments
  - SNOMED CT- provides a unified approach for most clinical answers (organisms, anatomic parts, specimens, diagnoses and symptoms) . It also provides codes for some observations .

THREE NLM SUPPORTED VOCABULARIES

Where to get

- SNOMED CT http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
- RX.NORM - http://www.nlm.nih.gov/research/umls/rxnorm/
  - Rx.Terms - https://wwwcf.nlm.nih.gov/umlslicense/rxtermApp/rxTerm.cfm
- LOINC - http://loinc.org/

LOINC

- LOINC Codes recommended by US federal government and other countries (Canada, Germany, China ,etc) for laboratory results and other content
- Also includes questionnaires (survey instruments as packages- with all the parts connected
  - E.g. PHQ-9 , PHQ-2, OASIS, MDS, CARE, etc
  - Working on PhenX variables and PRMISE
- RELMA DISC - Hand out Pig

SNOMED

- Also recommended widely and internationally
- More than 300K codes and hierarchical relations
- Has an elegant formalism

Rx.Norm

- US recommendation for drug ordering , medication profile, etc. Rx.Norm provides codes for drugs at the clinical drug and ingredient level. FDA provides related codes
- Clinical level includes the strength and dosage form
  - E.g. Ampicillin 500mg oral capsules
- Includes brand names and generic
- RX.Terms- a subset tailored to ease ordering (CMS)
  - AMIA 2008 Fall meeting Kin Wah Fung – paper
  - http://wwwcf.nlm.nih.gov/umlslicense/rxtermApp/rxTerm.cfm

Where to get

- ICD - http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
- LOINC - http://loinc.org/
- SNOMED CT http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
- RX.NORM - http://www.nlm.nih.gov/research/umls/rxnorm/
  - Rx.Terms - https://wwwcf.nlm.nih.gov/umlslicense/rxtermApp/rxTerm.cfm

Some general rules

- Favor measures over descriptors derived from the measures
- Separate special aspects of the measure separate questions to accommodate future changes
- Use “pick all that apply” format rather than “answer yes or no to each of item”.
- Be modest in collection goals (don’t try to capture everything

Favor measures over descriptors Hyperlipidemia as a lesson

- In the 1970’s a national data collection group measured cholesterol and recorded the trait - hyperlipidemia (yes/no)
- They did not keep the measurement
- What was the harm? The definition had crisp sharp edge: “Cholesterol > 300”

But…

- A few years later the criteria changed: "hypercholesterolemia" became "cholesterol > 250 mg" (It has changed again)
- So now way to directly compare trends in cholesterol
- Arrgh

The Lesson

- Definitions change - measures protect the future
- Compute the descriptors (from the measures when necessary)
- Accommodate new definitions by re-computing the descriptor

There is a place for categorical traits

- When the phenotype is categorical
  - E.g. Wingless (back to Drosophila)
  - Hemoglobins: S or SC or SS or CC
- When the primary source is categorical
  - E.g. ICD-9 discharge codes
  - Chart abstraction
- When time or funds are too short to allow a measurement
- Taxonomies – E.g. bacterial names, allergens

Where ontologies help

- Most help with task of systematic definitions for categorical things
- Realize they represent a mechanism for the entities and relationships within a defined world
- Less help in creating survey instrument where the questions are sentences (not descriptors)
- E.g. PHQ-9 In last week are you feeling bad about yourself, or that you are a failure or have let yourself or your family down

Ontology for phenotype descriptors

- Phenotype ontologies: the bridge between genomics and evolution
- Paula M. Mabee, Michael Ashburner, Quentin Cronk, Georgios V. Gkoutos, Melissa Haendel, Erik Segerdell, Chris Mungall and Monte Westerfield
- TRENDS in Ecology and Evolution Vol.22 No.7; 9 April 2007

Ontology More

- The Open Biomedical Ontologies (OBO) family of ontologies has 3 ontologies related to phenotypes:
- Mammalian phenotype ontology
  - pre-coordinated concepts (e.g., enlarged heart) -- used by the Jackson Laboratory to help researchers select particular strains of mice.
- http://www.bioontology.org/tools/portal/bioportal.html

Questions qualify the answers

- Questions about diagnoses/problems
  - Hospital discharge diagnoses (pulled from discharge summaries
  - Problems that are currently active
  - Major problems
  - Minor problems

CODE SYSTEMS

When creating new questions

- Recognize the many alternative styles for asking the questions
- We – the whole research field - should try to constrain these alternatives to reduce the number of variants

Alternative question asking, style 1

- Category vs. "continuous" answers
  - Smoking
    - Do you smoke? yes/no
    - How much do you smoke? (# packs per day)
    - A.M. time until first smoke
  - Renal failure
    - yes/no
    - versus last creatinine value
  - Cholesterol > 300: yes/no (earliest national survey)

One question - different answer lists

- Could tolerate different answer lists for different contexts – if they came from one universe

More Examples: Question/Answer Variation

Example from WHI

Education at screening?

- Didn't go to school
- Grade school (1-4 years)
- Grade school (5-8 years)
- Some high school (9-11 years)
- High school diploma /GED
- Vocational or training school
- Some college or Associate Degree
- College graduate or Baccalaureate Degree
- Some post-graduate or professional
- Master's Degree
- Doctoral Degree
- (PhD, M.D., J.D., etc.)

Example from Eye Study

1. What is the highest level of school you completed?

2. **Grade 11 or less**
3. **High school graduate**
4. **Some college or associate degree**
5. **Bachelor's degree**
6. **Postgraduate work**

Example Comparisons with WHI's Study

Example from WHI

F20 Current marital status

What is your current marital status?

(Mark the one that best describes you)

- Never Married
- Divorced or Separated
- Presently Married
- Widowed
- Marriage-like Relationship

Example from Eye Study

What is your current Marital Status

1. Never married
2. Divorced/separated
3. Widowed
4. Married

Pros and Cons 1

- Use numbers for continuous variables
- Numbers "always" better than broad categories when a physical scale exists
- Use years of schooling completed, number of packs of cigarettes per day
- Avoids variations due to answer lists
- But – list of descriptors might give be easier for patients to answer correctly

Alternative Styles 2

- Dichotomous questions about many states
  - COPD — yes, no (+,- many forms of negative)
  - CHF — yes, no
  - Stroke — yes, no
- Pick all that apply – (answer all that apply)
  - COPD [x]
  - CHF [ ]
  - Stroke [x]
  - None [ ]

Pros and Cons 2

- The first provides more information (reputedly)
- The second is easier and faster for the users
- Looks like one question to the user – I prefer it
- Should settle this question with empirical comparisons of user time cost

Alternatives: Specialized questions versus companion questions

- Specialized
  - Creatinine during last hospital stay = 3.1 mg/dl
  - Creatinine post cath = 3.1 mg/dl
- Generalized with companion variable
  - Creatinine = 3.1 mg/dl
  - Associated event = Post cath result

Pros and Cons 3

- Case 1- Better for data collector and analyzer
  - Prescribes what is to be entered
  - Easier for analysis of the given study
- Case 2 - better for standardization
  - Isolates differences; keeps commonalities across data sets
  - Facilitates data pooling
  - Provides direct linking to existing clinical care variables
- Solution – Can have both
  - Name the question as needed, then transform as needed for the study the two question for communication and pooling

Related question style issue — how to represent blood pressure

- Position related
  - Systolic BP Standing
  - Systolic BP Sitting
  - Systolic BP Lying
    - may be preferred for calculating standing/lying difference
- Site related
  - Systolic BP brachial
  - Systolic BP Radial
    - may be preferred for calculating brachial-ankle ratio

Style more

- But there can be more to measuring BP
  - Cuff size
  - Method
  - Alternative locations
  - Relation to exercise

Style still more

- So define as a panel – with many optional elements
- Diastolic BP
- Systolic BP
  - Cuff Size
  - BP method (auscultatory manual, auscultatory auto, oscillometery, etc)
  - BP vendor and model name (esp when delivered automatically
  - BP Serial number (when delivered automatically
  - Always want time stamped
  - Who took (maybe)
  - Where measured (maybe) – e.g. home/office/hospital

Two step vs. one step question

- From STS – Two step
  - Angina? yes /no If yes:
  - Angina type? – stable/unstable
- Versus - one step
  - Angina? None/stable/unstable
- From STS – one step
  - Radial artery used? No, radial/left/right/both
- When one-answer can be part of next question it saves a separate user response, and removes a source of differences between questions

What's it for

- What questions are being asked and what analysis is planned
- Makes a difference in what you collect and how much
- Realize the deep versus wide conundrum
  - If you collect hoards of variables – you need even larger hoards of patients for analyses
  - There are trade offs
  - Fewer variables on more patients is usually a better bet.